Variational Inference for the Indian Buffet Process

Finale Doshi-Velez* Engineering Department Cambridge University Cambridge, UK

Kurt T. Miller* Computer Science Division University of California, Berkeley Berkeley, CA

Jurgen Van Gael* Engineering Department Cambridge University Cambridge, UK

Yee Whye Teh Gatsby Unit University College London London, UK

Abstract

The Indian Buffet Process (IBP) is a nonparametric prior for latent feature models in which observations are influenced by a combination of hidden features. For example, images may be composed of several objects and sounds may consist of several notes. Latent feature models seek to infer these unobserved features from a set of observations; the IBP provides a principled prior in situations where the number of hidden features is unknown. Current inference methods for the IBP have all relied on sampling. While these methods are guaranteed to be accurate in the limit, samplers for the IBP tend to mix slowly in practice. We develop a deterministic variational method for inference in the IBP based on a truncated stick-breaking approximation, provide theoretical bounds on the truncation error, and evaluate our method in several data regimes.

1 INTRODUCTION

Many unsupervised learning problems seek to identify a set of unobserved, co-occurring features from a set of observations. For example, given images composed of various objects, we may wish to identify the set of unique objects and determine which images contain which objects. Similarly, we may wish to extract a set of notes or chords from an audio file as well as when each note was played. In scenarios such as these, the number of latent features is often unknown a priori.

*Authors contributed equally

Appearing in Proceedings of the 12^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2009, Clearwater Beach, Florida, USA. Volume 5 of JMLR: W&CP 5.

Unfortunately, even though the true number of features is unknown, most traditional machine learning approaches take the number of latent features as an input. In these situations, standard model selection approaches define and manage the trade-off between model complexity and model fit. In contrast, nonparametric Bayesian approaches treat the number of features as a random quantity to be determined as part of the posterior inference procedure.

The most common nonparametric prior for latent feature models is the Indian Buffet Process (IBP) (Griffiths & Ghahramani, 2005). The IBP is a prior on infinite binary matrices that allows us to simultaneously infer which features influence a set of observations and how many features there are. The form of the prior ensures that only a finite number of features will be present in any finite set of observations, but more features may appear as more observations are received. This property is both natural and desirable if we consider, for example, a set of images: any one image contains a finite number of objects, but, as we see more images, we expect to see objects not present in the previous images.

While an attractive model, the combinatorial nature of the IBP makes inference particularly challenging. Even if we limit ourselves to K features for N objects, there exist $O(2^{NK})$ possible feature assignments. As a result, sampling-based inference procedures for the IBP often suffer because they assign specific values to the feature-assignment variables. Hard variable assignments give samplers less flexibility to move between optima, and the samplers may need large amounts of time to escape small optima and find regions with high probability mass. Unfortunately, all current inference procedures for the IBP rely on sampling. These approaches include Gibbs sampling (Griffiths & Ghahramani, 2005) (which may be augmented with Metropolis split-merge proposals (Meeds et al., 2007)), slice sampling (Teh et al., 2007), and particle filtering (Wood & Griffiths, 2007).

Mean field variational methods, which approximate the true posterior via a simpler distribution, provide a deterministic alternative to sampling-based approaches. Inference involves using optimisation techniques to find a good approximate posterior. For the IBP, the approximating distribution maintains a separate probability for each feature-observation assignment. Optimising these probability values is also fraught with local optima, but the soft variable assignments give the variational method flexibility lacking in the samplers. In the early stages of the inference, the soft-assignments can help the variational method avoid bad local optima. Several variational approximations have provided benefits for other nonparametric Bayesian models, including Dirichlet Processes (e.g. (Blei & Jordan, 2004)) and Gaussian Processes (e.g. (Winther, 2000)). Of all the nonparametric Bayesian models studied so far, however, the IBP is the most combinatorial and is therefore in the most need of a more efficient inference algorithm.

The rest of the paper is organised as follows. Section 2 reviews the IBP model and current sampling-based inference techniques. Section 3 presents our variational approach based on a truncated representation of the IBP. Building on ideas from (Teh et al., 2007) and (Thibaux & Jordan, 2007), we also derive bounds on the expected error due to the use of a truncated approximation; these bounds can serve as guidelines for what level of truncation may be appropriate. Section 4 demonstrates how our variational approach allows us to scale to higher dimensional data sets while still getting good predictive results.

2 THE INDIAN BUFFET PROCESS

Let X be an $N \times D$ matrix where each of the N rows contains a D-dimensional observation. In this paper, we focus on a model in which X can be approximated by ZA where Z is an $N \times K$ binary matrix and A is a $K \times D$ matrix. Each column of Z corresponds to the presence of a latent feature; $z_{nk} = Z(n,k)$ is one if feature k is present in observation n and zero otherwise. The values for feature k are stored in row k of A. The observed data X is then given by $ZA + \epsilon$, where ϵ is some measurement noise (see Figure 1). We assume that the noise is independent of Z and A and is uncorrelated across observations.

Given X, we wish to find the posterior distribution of Z and A. We do this using Bayes rule

$$p(Z, A|X) \propto p(X|Z, A)p(Z)p(A)$$

where we have assumed that Z and A are a priori independent. The application will determine the likelihood function $p(X|Z, A)$ and the feature prior $p(A)$. We are left with placing a prior on Z. Since we often do not know K, we wish to place a flexible prior on Z that allows K to be determined at inference time.

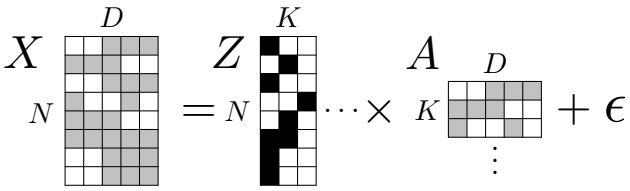

Figure 1: The latent feature model proposes the data X is the product of Z and A with some noise.

2.1 THE IBP PRIOR

The Indian Buffet Process places the following prior on $[Z]$, a canonical form of Z that is invariant to the ordering of the features (see (Griffiths & Ghahramani, 2005) for details):

$$p([Z]) = \frac{\alpha^K}{\prod_{h \in \{0,1\}^N \setminus \mathbf{0}} K_h!} \exp\{-\alpha H_N\} \cdot \prod_{k=1}^{K} \frac{(N - m_k)!(m_k - 1)!}{N!}, \quad (1)$$

where K is the number of nonzero columns in Z, m_k is the number of ones in column k of Z, H_N is the N^{th} harmonic number, and K_h is the number of occurrences of the non-zero binary vector h among the columns in Z. The parameter α controls the expected number of features present in each observation.

The following culinary metaphor is one way to sample a matrix Z from the prior described in Equation (1). Imagine the rows of Z correspond to customers and the columns correspond to dishes in an infinitely long (Indian) buffet. The first customer takes the first $\mathsf{Poisson}(\alpha)$ dishes. The i^{th} customer then takes dishes that have been previously sampled with probability m_k/i, where m_k is the number of people who have already sampled dish k. He also takes $\mathsf{Poisson}(\alpha/i)$ new dishes. Then, z_{nk} is one if customer n tried the k^{th} dish and zero otherwise. This process is infinitely exchangeable, which means that the order in which the customers attend the buffet has no impact on the distribution of Z (up to permutations of the columns).

The Indian buffet metaphor leads directly to a Gibbs sampler. Bayes' rule states $p(z_{nk}|Z_{-nk}, A, X) \propto p(X|A, Z)p(z_{nk}|Z_{-nk})$. The likelihood term $p(X|A, Z)$ is easily computed from the noise model while the prior term $p(z_{nk}|Z_{-nk})$ is obtained by assuming that customer n was the last to enter the restaurant (this assumption is valid due to exchangeability). The prior

term is $p(z_{nk}|Z_{-nk}) = m_K/N$ for active features. New features are sampled by combining the likelihood model with the $\mathsf{Poisson}(\alpha/N)$ prior on the number of new dishes a customer will try. When the prior on A is conjugate to the likelihood model, A can be marginalised out, resulting in a collapsed Gibbs sampler. If the likelihood is not conjugate, or, as in the linear-Gaussian model, if $p(X|Z)$ is much more expensive to compute than $p(X|Z, A)$, we can also sample the matrix A based on its posterior distribution.

2.2 STICK-BREAKING CONSTRUCTION

While the restaurant construction of the IBP directly lends itself to a Gibbs sampler, the stick-breaking construction of (Teh et al., 2007) is at the heart of our variational approach. To generate a matrix Z from the IBP prior using the stick-breaking construction, we begin by assigning a parameter $\pi_k \in (0, 1)$ to each column of Z. Given π_k, each z_{nk} in column k is sampled as an independent $\mathsf{Bernoulli}(\pi_k)$. Since each 'customer' samples a dish independently of the other customers, it is clear in this representation that the ordering of the customers does not impact the distribution.

The π_k themselves are generated by a stick-breaking process. We first draw a sequence of independent random variables $v_1, v_2, \ldots$, each distributed $\mathsf{Beta}(\alpha, 1)$. Next, we let $\pi_1 = v_1$. For each subsequent k, we let $\pi_k = v_k\pi_{k-1} = \prod_{i=1}^{k} v_i$, resulting in a decreasing sequence of weights π_k. The expression for π_k shows that, in a set of N observations, the probability of seeing feature k decreases exponentially with k. We also see that larger values of α mean that we expect to see more features in the data.

3 VARIATIONAL INFERENCE

In this section, we focus on variational inference procedures for the linear-Gaussian likelihood model (Griffiths & Ghahramani, 2005), in which A and ϵ are zero mean Gaussians with variances σ_A^2 and σ_n^2 respectively. However, the updates can be adapted to other exponential family likelihood models. As an example, we briefly discuss the infinite ICA model (Knowles & Ghahramani, 2007).

We denote the set of hidden variables in the IBP by $\boldsymbol{W} = \{\boldsymbol{\pi}, \boldsymbol{Z}, \boldsymbol{A}\}$ and the set of parameters by $\boldsymbol{\theta} = \{\alpha, \sigma_A^2, \sigma_n^2\}$. Computing the true log posterior $\ln p(\boldsymbol{W}|\boldsymbol{X}, \boldsymbol{\theta}) = \ln p(\boldsymbol{W}, \boldsymbol{X}|\boldsymbol{\theta}) - \ln p(\boldsymbol{X}|\boldsymbol{\theta})$ is difficult due to the intractability of computing the log marginal probability $\ln p(\boldsymbol{X}|\boldsymbol{\theta}) = \ln \int p(\boldsymbol{X}, \boldsymbol{W}|\boldsymbol{\theta}) d\boldsymbol{W}$.

Mean field variational methods approximate the true posterior with a *variational distribution* $q(\boldsymbol{W})$ from some tractable family of distributions Q (Beal, 2003; Wainwright & Jordan, 2008). Inference in this approach then reduces to finding the member $q \in Q$ that minimises the KL divergence $D(q(\boldsymbol{W})||p(\boldsymbol{W}|\boldsymbol{X}, \boldsymbol{\theta}))$. Since the KL divergence $D(q||p)$ is nonnegative and equal to zero iff $p = q$, the unrestricted solution to our problem is to set $q(\boldsymbol{W}) = p(\boldsymbol{W}|\boldsymbol{X}, \boldsymbol{\theta})$. However, this general optimisation problem is intractable. We therefore restrict Q to a parameterised family of distributions for which this optimisation is tractable. For the IBP, we will let Q be the factorised family

$$q(\boldsymbol{W}) = q_{\boldsymbol{\tau}}(\boldsymbol{\pi}) q_{\boldsymbol{\phi}}(\boldsymbol{A}) q_{\boldsymbol{\nu}}(\boldsymbol{Z}) \tag{2}$$

where $\boldsymbol{\tau}$, $\boldsymbol{\phi}$, and $\boldsymbol{\nu}$ are the variational parameters that we optimise to minimise $D(q||p)$. Inference then consists of optimising the parameters of the approximating distribution to most closely match the true posterior. This optimisation is equivalent to maximising a lower bound on the evidence:

$$\begin{aligned} & \arg\max_{\tau,\phi,\nu} \ln p(\boldsymbol{X}|\boldsymbol{\theta}) - D(q||p) \\ = \; & \arg\max_{\tau,\phi,\nu} H[q] + \mathbb{E}_q[\ln(p(\boldsymbol{X}, \boldsymbol{W}|\boldsymbol{\theta})]. \end{aligned} \tag{3}$$

where $H[q]$ is the entropy of distribution q. Therefore, to minimise $D(q||p)$, we can iteratively update the variational parameters so as to maximise the right side of Equation (3).

We derive two mean field approximations, both of which apply a truncation level K to the maximum number of features in the variational distribution. The first minimises the KL-divergence between the variational distribution and a finite approximation p_K to the IBP described below; we refer to this approach as the *finite variational* method. The second approach minimises the KL-divergence to the true IBP posterior. We call this approach the *infinite variational* method because, while our variational distribution is finite, its updates are based the true IBP posterior over an infinite number of features.

Most of the required expectations are straightforward to compute, and many of the parameter updates follow directly from standard update equations for variational inference in the exponential family (Beal, 2003; Wainwright & Jordan, 2008). We focus on the nontrivial computations and reserve the full update equations for a technical report.

3.1 FINITE VARIATIONAL APPROACH

The finite variational method uses a finite Beta-Bernoulli approximation to the IBP (Griffiths & Ghahramani, 2005). The finite Beta-Bernoulli model with K features first draws each feature's probability π_k independently from $\mathsf{Beta}(\alpha/K, 1)$. Then, each z_{nk} is independently drawn from $\mathsf{Bernoulli}(\pi_k)$ for all n.

Our finite variational approach approximates the true IBP model $p(\boldsymbol{W}, \boldsymbol{X}|\boldsymbol{\theta})$ with $p_K(\boldsymbol{W}, \boldsymbol{X}|\boldsymbol{\theta})$ in Equation (3) where p_K uses the prior on Z defined by the finite Beta-Bernoulli model. While variational inference in the finite Beta-Bernoulli model is not the same as variational inference with respect to the true IBP posterior, the variational updates are significantly more straightforward and, in the limit of large K, the finite Beta-Bernoulli approximation is equivalent to the IBP. We use a fully factorised variational distribution $q_{\boldsymbol{\tau}_k}(\pi_k) = \mathsf{Beta}(\pi_k; \tau_{k1}, \tau_{k2})$, $q_{\boldsymbol{\phi}_k}(\boldsymbol{A}_{k\cdot}) = \mathsf{Normal}(\boldsymbol{A}_{k\cdot}; \bar{\boldsymbol{\phi}}_k, \boldsymbol{\Phi}_k)$, $q_{\nu_{nk}}(z_{nk}) = \mathsf{Bernoulli}(z_{nk}; \nu_{nk})$.

3.2 INFINITE VARIATIONAL APPROACH

The second variational approach, similar to the one used in (Blei & Jordan, 2004), uses a truncated version of the stick-breaking construction for the IBP as the approximating variational model q. Instead of directly approximating the distribution of π_k in our variational model, we will work with the distribution of the stick-breaking variables $\boldsymbol{v} = \{v_1, \ldots, v_K\}$. In our truncated model with truncation level K, the probability π_k of feature k is $\prod_{i=1}^{k} v_i$ for $k \leq K$ and zero otherwise. The advantage of using $\boldsymbol{v}$ as our hidden variable is that under the IBP prior, the $\{v_1 \ldots v_K\}$ are independent draws from the Beta distribution, whereas the $\{\pi_1 \ldots \pi_K\}$ are dependent. We therefore use the factorised variational distribution $q(\boldsymbol{W}) = q_{\boldsymbol{\tau}}(\boldsymbol{v}) q_{\boldsymbol{\phi}}(\boldsymbol{A}) q_{\boldsymbol{\nu}}(\boldsymbol{Z})$ where $q_{\boldsymbol{\tau}_k}(v_k) = \mathsf{Beta}(v_k; \tau_{k1}, \tau_{k2})$, $q_{\boldsymbol{\phi}_k}(\boldsymbol{A}_{k\cdot}) = \mathsf{Normal}(\boldsymbol{A}_{k\cdot}; \bar{\boldsymbol{\phi}}_k, \Phi_k)$, and $q_{\nu_{nk}}(z_{nk}) = \mathsf{Bernoulli}(Z_{nk}; \nu_{nk})$.

3.3 VARIATIONAL LOWER BOUND

We split the expectation in Equation (3) into terms depending on each of the latent variables. Here, $\boldsymbol{v}$ are the stick-breaking parameters in the infinite approach; the expression for the finite Beta approximation is identical except with $\boldsymbol{\pi}$ substituted into the expectations.

$$\begin{aligned}\ln p(\boldsymbol{X}|\boldsymbol{\theta}) \geq & H[q] + \textstyle\sum_{k=1}^{K} \mathbb{E}_{\boldsymbol{v}} \left[\ln p(v_k|\alpha)\right] \\ & + \textstyle\sum_{k=1}^{K} \mathbb{E}_{\boldsymbol{A}} \left[\ln p(\boldsymbol{A}_{k\cdot}|\sigma_A^2)\right] \\ & + \textstyle\sum_{k=1}^{K} \sum_{n=1}^{N} \mathbb{E}_{\boldsymbol{v},\boldsymbol{Z}} \left[\ln p(z_{nk}|\boldsymbol{v})\right] \\ & + \textstyle\sum_{n=1}^{N} \mathbb{E}_{\boldsymbol{Z},\boldsymbol{A}} \left[\ln p(\boldsymbol{X}_{n\cdot}|\boldsymbol{Z}, \boldsymbol{A}, \sigma_n^2)\right]\end{aligned}$$

In the finite Beta approximation, all of the expectations are straightforward exponential family calculations. In the infinite case, the key difficulty lies in computing the expectations $\mathbb{E}_{\boldsymbol{v},\boldsymbol{Z}} \left[\ln p(z_{nk}|\boldsymbol{v})\right]$. We decompose this expectation as

$$\begin{aligned}& \mathbb{E}_{\boldsymbol{v},\boldsymbol{Z}} \left[\ln p(z_{nk}|\boldsymbol{v})\right] \\ =& \mathbb{E}_{\boldsymbol{v},\boldsymbol{Z}} \left[\ln p(z_{nk} = 1|\boldsymbol{v})^{\mathbb{I}(z_{nk}=1)} p(z_{nk} = 0|\boldsymbol{v})^{\mathbb{I}(z_{nk}=0)}\right] \\ =& \nu_{nk} \left(\textstyle\sum_{m=1}^{k} \psi(\tau_{k2}) - \psi(\tau_{k1} + \tau_{k2})\right) \\ & + (1 - \nu_{nk}) \mathbb{E}_{\boldsymbol{v}} \left[\ln \left(1 - \textstyle\prod_{m=1}^{k} v_m\right)\right]\end{aligned}$$

where $\mathbb{I}(\cdot)$ is the indicator function that its argument is true and $\psi(\cdot)$ is the digamma function. We are still left with the problem of evaluating the expectation $\mathbb{E}_{\boldsymbol{v}}[\ln(1 - \prod_{m=1}^{k} v_m)]$, or alternatively, computing a lower bound for the expression.

There are computationally intensive methods for finding arbitrarily good lower bounds for this term using a Taylor series expansion of $\ln(1 - x)$. However, we present a more computationally efficient bound that is only slightly looser. We first introduce a multinomial distribution $q_k(y)$ that we will optimise to get as tight a lower bound as possible and use Jensen's inequality:

$$\begin{aligned}& \mathbb{E}_{\boldsymbol{v}} \left[\ln \left(1 - \textstyle\prod_{m=1}^{k} v_m\right)\right] \\ =& \mathbb{E}_{\boldsymbol{v}} \left[\ln \left(\textstyle\sum_{y=1}^{k} q_k(y) \frac{(1 - v_y) \prod_{m=1}^{y-1} v_m}{q_k(y)}\right)\right] \\ \geq & \mathbb{E}_{\boldsymbol{v}} \mathbb{E}_y \left[\ln \left((1 - v_y) \textstyle\prod_{m=1}^{y-1} v_m\right) - \ln q_k(y)\right] \\ =& \mathbb{E}_y \left[\psi(\tau_{y2}) + \textstyle\sum_{m=1}^{y-1} \psi(\tau_{m1}) - \sum_{m=1}^{y} \psi(\tau_{m1} + \tau_{m2})\right] \\ & + H(q_k).\end{aligned}$$

These equations hold for any q_k. We take derivatives to find the q_k that maximises the lower bound:

$$q_k(y) \propto e^{\left(\psi(\tau_{y2}) + \sum_{m=1}^{y-1} \psi(\tau_{m1}) - \sum_{m=1}^{y} \psi(\tau_{m1} + \tau_{m2})\right)}$$

where the proportionality is required to make q_k a valid distribution. We can plug this multinomial lower bound for $\mathbb{E}_{\boldsymbol{v},\boldsymbol{Z}} \left[\ln p(z_{nk}|\boldsymbol{v})\right]$ back into the lower bound on $\ln p(\boldsymbol{X}|\boldsymbol{\theta})$ and then optimise this lower bound.

3.4 PARAMETER UPDATES

The parameter updates in the finite model are all straightforward updates from the exponential family (Wainwright & Jordan, 2008). In the infinite case, updates for the variational parameter for $\boldsymbol{A}$ remain standard exponential family updates. The update on $\boldsymbol{Z}$ is also relatively straightforward to compute

$$\begin{aligned}q_{\nu_{nk}}(z_{nk}) \propto & \exp\left(\mathbb{E}_{\boldsymbol{v},\boldsymbol{A},\boldsymbol{Z}_{-nk}} \left[\ln p(\boldsymbol{W}, \boldsymbol{X}|\boldsymbol{\theta})\right]\right) \\ \propto & \exp\Big(\mathbb{E}_{\boldsymbol{A},\boldsymbol{Z}_{-nk}} (\ln p(X_{n\cdot}|\boldsymbol{Z}_{n\cdot}, \boldsymbol{A}, \sigma_n^2)) \\ & + \mathbb{E}_{\boldsymbol{v}} (\ln p(z_{nk}|\boldsymbol{v}))\Big),\end{aligned}$$

where we can again approximate $\mathbb{E}_{\boldsymbol{v}}(\ln p(z_{nk}|\boldsymbol{v}))$ with a Taylor series or the multinomial method presented in Section 3.3.

The update for the stick-breaking variables $\boldsymbol{v}$ is more complex because the variational updates no longer stay in the exponential family due to the terms $\mathbb{E}_{\boldsymbol{v}}(\ln p(z_{nk}|\boldsymbol{v}))$. If we use a Taylor series approximation for this term, we no longer have closed form updates for $\boldsymbol{v}$ and must resort to numerical optimisation. If we use the multinomial lower bound, then for fixed $q_k(y)$, terms decompose independently for each v_m and we get a closed form exponential family update. We will use the latter approach in our results section.

3.5 ICA MODEL

An infinite version of the ICA model based on the IBP was introduced by (Knowles & Ghahramani, 2007). Instead of simply modeling the data X as centered around ZA, the infinite ICA model introduces an additional signal matrix S so that the data is centered around $(Z \odot S)A$, where $\odot$ denotes element-wise multiplication. Placing a Laplace prior on S allows it to modulate the feature assignment matrix Z. Variational updates for the infinite ICA model are straightforward except those for S: we apply a Laplace approximation and numerically optimise the parameters.

3.6 TRUNCATION ERROR

Both of our variational inference approaches require us to choose a truncation level K for our variational distribution. Building on results from (Thibaux & Jordan, 2007; Teh et al., 2007), we present bounds on how close the marginal distributions are when using a truncated stick-breaking prior and the true IBP stick-breaking prior. Our development parallels bounds for the Dirichlet Process by (Ishwaran & James, 2001) and presents the first such truncation bounds for the IBP.

Intuitively, the error in the truncation will depend on the probability that, given N observations, we observe features beyond the first K in the data (otherwise the truncation should have no effect). Let us denote the marginal distribution of observation X by $m_\infty(X)$ when we integrate over W drawn from the IBP. Let $m_K(X)$ be the marginal distribution when W are drawn from the truncated stick-breaking prior with truncation level K.

Using the Beta Process representation for the IBP (Thibaux & Jordan, 2007) and using an analysis similar to the one in (Ishwaran & James, 2001), we can show that the difference between these distributions is at most

$$\begin{aligned}
&\frac{1}{4}\int |m_K(X) - m_\infty(X)|dX \\
&\quad\leq \Pr(\exists k > K, n \text{ with } z_{nk} = 1) \\
&\quad= 1 - \Pr\left(\text{all } z_{ik} = 0,\ i \in \{1,\ldots,N\}, k > K\right) \\
&\quad= 1 - \mathbb{E}\left[\left(\prod_{i=K+1}^{\infty}(1-\pi_i)\right)^N\right] \qquad (4)
\end{aligned}$$

We present here one formal bound for this difference. The extended version of this paper will include similar bounds which can be derived directly by applying Jensen's inequality to the expectation above as well as a heuristic bound which tends to be tighter in practice.

We begin the derivation of the truncation bound by applying Jensen's inequality to equation (4):

$$-\mathbb{E}\left[\left(\prod_{i=K+1}^{\infty}(1-\pi_i)\right)^N\right] \leq -\left(\mathbb{E}\left[\prod_{i=K+1}^{\infty}(1-\pi_i)\right]\right)^N \qquad (5)$$

The Beta Process construction for the IBP (Thibaux & Jordan, 2007) implies that the sequence $\pi_1, \pi_2, \ldots$ can be modeled as a Poisson process on the unit interval $(0, 1)$ with rate $\mu(x)dx = \alpha x^{-1}dx$. It follows that the unordered truncated sequence $\pi_{K+1}, \pi_{K+2}, \ldots$ may be modeled as a Poisson process on the interval $(0, \pi_K)$ with the same rate. The Levy-Khintchine formula states that the moment generating function of a Poisson process X with rate μ can be written as

$$\mathbb{E}[\exp(f(X))] = \exp\left(\int(\exp(f(x)) - 1)\mu(x)dx\right)$$

where $f(X) = \sum_{x\in X} f(x)$. We apply the Levy-Khintchine formula to simplify the inner expectation of equation (5):

$$\begin{aligned}
\mathbb{E}\left[\prod_{i=K+1}^{\infty}(1-\pi_i)\right] &= \mathbb{E}\left[\exp\left(\sum_{i=K+1}^{\infty}\ln(1-\pi_i)\right)\right] \\
&= \mathbb{E}_{\pi_K}\left[\exp\left(\int_0^{\pi_K}(\exp(\ln(1-x)) - 1)\,\mu(x)dx\right)\right] \\
&= \mathbb{E}_{\pi_K}\left[\exp\left(-\alpha\pi_K\right)\right].
\end{aligned}$$

Finally, we apply Jensen's inequality, using the fact that π_K is the product of independent $\mathsf{Beta}(\alpha, 1)$ variables to get

$$\begin{aligned}
\mathbb{E}_{\pi_K}[\exp\left(-\alpha\pi_K\right)] &\geq \exp\left(\mathbb{E}_{\pi_k}[-\alpha\pi_K]\right) \\
&= \exp\left(-\alpha\left(\frac{\alpha}{1+\alpha}\right)^K\right)
\end{aligned}$$

Substituting the expression into equation (5) gives

$$\frac{1}{4}\int |m_K(X) - m_\infty(X)|dX \leq 1 - \exp\left(-N\alpha\left(\frac{\alpha}{1+\alpha}\right)^K\right) \qquad (6)$$

Similar to truncation bound for the Dirichlet Process, we see that for fixed K, the expected error increases with N and α— the factors that increase the expected number of features in a dataset. However, the bound decreases exponentially quickly as K is increased.

Figure 2 shows our truncation bound and the true L_1 distance based on 1000 Monte Carlo simulations of an IBP matrix with $N = 30$ observations and $\alpha = 5$. As expected, the bound decreases exponentially fast with the truncation level K. However, the bound is fairly loose. In practice, we find that heuristic bound using Taylor expansions (see extended version) provides much tighter estimates of the loss.

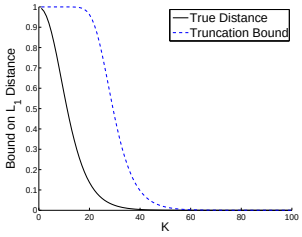

Figure 2: Truncation bound and true L_1 distance.

4 RESULTS

We compared our variational approaches with both Gibbs sampling and particle filtering. Mean field variational algorithms are only guaranteed to converge to a *local* optima, so we applied standard optimisation tricks to avoid issues of bad minima. Each run was given a number of random restarts and the hyperparameters for the noise and feature variance were tempered to smooth the posterior. We also experimented with several other techniques such as gradually introducing data and merging correlated features that were less useful as the size and dimensionality of the datasets increased; they were not included in the final experiments.

The sampling methods we compared against were the collapsed Gibbs sampler of (Griffiths & Ghahramani, 2005) and a partially-uncollapsed alternative in which instantiated features are explicitly represented and new features are integrated out. In contrast to the variational methods, the number of features present in the IBP matrix will adaptively grow or shrink in the samplers. To provide a fair comparison with the variational approaches, we also tested finite variants of the collapsed and uncollapsed Gibbs samplers. Finally, we also tested against the particle filter of (Wood & Griffiths, 2007). All sampling methods were tempered and given an equal number of restarts as the variational methods.

Both the variational and Gibbs sampling algorithms were heavily optimised for efficient matrix computation so we could evaluate the algorithms both on their running times and the quality of the inference. For the particle filter, we used the implementation provided by (Wood & Griffiths, 2007). To measure the quality of these methods, we held out one third of the observations on the last half of the dataset. Once the inference was complete, we computed the predictive likelihood of the held out data and averaged over restarts.

4.1 SYNTHETIC DATA

The synthetic datasets consisted of Z and A matrices randomly generated from the truncated stick-breaking prior. Figure 3 shows the evolution of the test-likelihood over a thirty minute interval for a dataset with 500 observations of 500 dimensions each generated with 20 latent features.[1] The error bars indicate the variation over the 5 random starts. The finite uncollapsed Gibbs sampler (dotted green) rises quickly but consistently gets caught in a lower optima and has higher variance. This variance is not due to the samplers mixing, but instead due to each sampler getting stuck in widely varying local optima. The variational methods are slightly slower per iteration but soon find regions of higher predictive likelihoods. The remaining samplers are much slower per iteration, often failing to mix within the allotted interval.

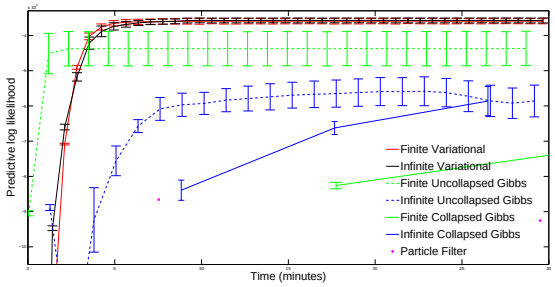

Figure 3: Evolution of test log-likelihoods over a thirty-minute interval for $N = 500$, $D = 500$, and $K = 20$. The finite uncollapsed Gibbs sampler has the fastest rise but gets caught in a lower optima than the variational approach.

Figures 4 and 5 show results from a systematic series of tests in which we tested all combinations of observation count $N = \{5, 10, 50, 100, 500, 1000\}$, dimensionality $D = \{5, 10, 50, 100, 500, 1000\}$, and truncation

[1]The particle filter must be run to completion before making prediction, so we cannot test its predictive performance over time. We instead plot the test likelihood only at the end of the inference for particle filters with 10 and 50 particles (the two magenta points).

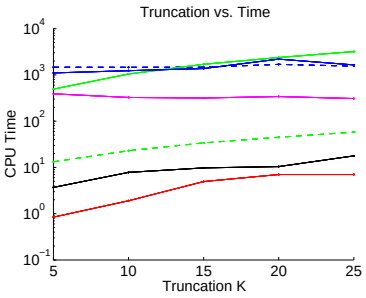

Figure 4: Time versus truncation (K). The variational approaches are generally orders of magnitude faster than the samplers (note log scale on the time axis).

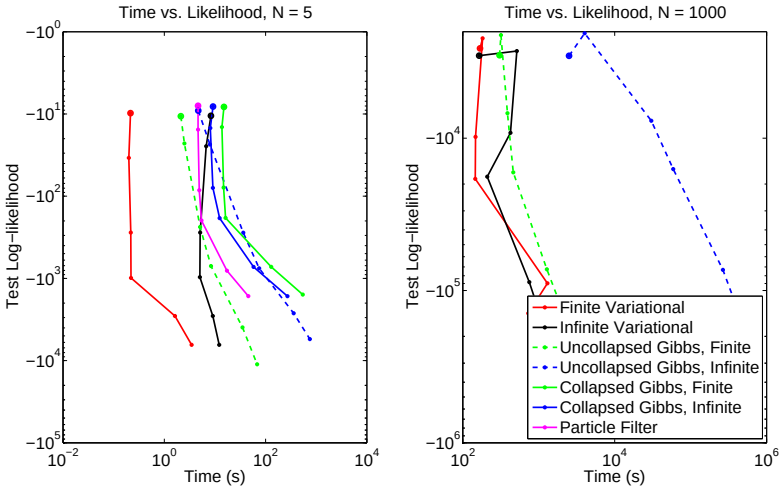

Figure 5: Time versus log-likelihood plot for $K = 20$. The larger dots correspond to $D = 5$ the smaller dots to $D = 10, 50, 100, 500, 1000$.

level $K = \{5, 10, 15, 20, 25\}$. Each of the samplers was run for 1000 iterations on three chains and the particle filter was run with 500 particles. For the variational methods, we used a stopping criterion that halted the optimisation when the variational lower bound between the current and previous iterations changed by a multiplicative factor of less than 10^{-4} and the tempering process had completed.

Figure 4 shows how the computation time scales with the truncation level. The variational approaches and the uncollapsed Gibbs are consistently an order of magnitude faster than other algorithms.

Figure 5 shows the interplay between dimensionality, computation time, and test log-likelihood for datasets of size $N = 5$ and $N = 1000$ respectively. For $N = 1000$, the collapsed Gibbs samplers and particle filter did not finish, so they do not appear on the plot. We chose $K = 20$ as a representative truncation level. Each line represents increasing dimensionality for a particular method (the large dot indicates $D = 5$, the subsequent dots correspond to $D = 10$, 50, etc.). The nearly vertical lines of the variational methods show that they are quite robust to increasing dimension. As dimensionality and dataset size increase, the variational methods become increasingly faster than the samplers. By comparing the lines across the likelihood dimension, we see that for the very small dataset, the variational method often has a lower test log-likelihood than the samplers. In this regime, the samplers are fast to mix and explore the posterior. However, the test log-likelihoods are comparable for the larger dataset.

4.2 REAL DATA

We next tested two real-world datasets to show how our approach fared with complex, noisy data not drawn from the IBP prior (our main goal was not to demonstrate low-rank approximations). The Yale Faces (Georghiades et al., 2001) dataset consisted of 721 32x32 pixel frontal-face images of 14 people with varying expressions and lighting conditions. We set σ_a and σ_n based on the variance of the data. The speech dataset consisted of 245 observations sampled from a 10-microphone audio recording of 5 different speakers. We applied the ICA version of our inference algorithm, where the mixing matrix S modulated the effect of each speaker on the audio signals. The feature and noise variances were taken from an initial run of the Gibbs sampler where σ_n and σ_a were also sampled.

Tables 1 and 2 show the results for each of the datasets. All Gibbs samplers were uncollapsed and run for 200 iterations.[2] In the higher dimensional Yale dataset, the variational methods outperformed the uncollapsed Gibbs sampler. When started from a random position, the uncollapsed Gibbs sampler quickly became stuck in a local optima. The variational method was able to find better local optima because it was initially very uncertain about which features were present in which data points; expressing this uncertainty explicitly through the variational parameters (instead of through a sequence of samples) allowed it the flexibility to improve upon its bad initial starting point.

[2]On the Yale dataset, we did not test the collapsed samplers because the finite collapsed Gibbs sampler required one hour per iteration with $K = 5$ and the infinite collapsed Gibbs sampler generated one sample every 50 hours. In the iICA model, the features $\boldsymbol{A}$ cannot be marginalised.

The story for the speech dataset, however, is quite different. Here, the variational methods were not only slower than the samplers, but they also achieved lower test-likelihoods. The evaluation on the synthetic datasets points to a potential reason for the difference: the speech dataset is much simpler than the Yale dataset, consisting of 10 dimensions (vs. 1032 in the Yale dataset). In this regime, the Gibbs samplers perform well and the approximations made by the variational method become apparent. As the dimensionality grows, the samplers have more trouble mixing, but the variational methods are still able to find regions of high probability mass.

Table 1: Running times in seconds and test log-likelihoods for the Yale Faces dataset.

Algorithm	**K**	**Time**	**Test Log-Likelihood** ($\times 10^6$)
Finite Gibbs	5	464.19	-2.250
	10	940.47	-2.246
	25	2973.7	-2.247
Finite Variational	5	163.24	-1.066
	10	767.1	-0.908
	25	10072	-0.746
Infinite Variational	5	176.62	-1.051
	10	632.53	-0.914
	25	19061	-0.750

Table 2: Running times in seconds and test log-likelihoods for the speech dataset.

Algorithm	**K**	**Time**	**Test Log-Likelihood**
Finite Gibbs	2	56	-0.7444
	5	120	-0.4220
	9	201	-0.4205
Infinite Gibbs	na	186	-0.4257
Finite Variational	2	2477	-0.8455
	5	8129	-0.5082
	9	8539	-0.4551
Infinite Variational	2	2702	-0.8810
	5	6065	-0.5000
	9	8491	-0.5486

5 SUMMARY

The combinatorial nature of the Indian Buffet Process poses specific challenges for sampling-based inference procedures. In this paper, we derived a mean field variational inference procedure for the IBP. Whereas sampling methods work in the discrete space of binary matrices, the variational method allows for soft assignments of features because it approaches the inference problem as a continuous optimisation. We showed experimentally that, especially for high dimensional problems, the soft assignments allow the variational methods to explore the posterior space faster than sampling-based approaches.

Acknowledgments

FD was supported by a Marshall scholarship. KTM was supported by contract DE-AC52-07NA27344 from the U.S. Department of Energy through Lawrence Livermore National Laboratory. JVG was supported by a Microsoft Research scholarship.

References

Beal, M. J. (2003). *Variational algorithms for approximate bayesian inference*. Doctoral dissertation, Gatsby Computational Neuroscience Unit, UCL.

Blei, D., & Jordan, M. (2004). Variational methods for the Dirichlet process. *Proceedings of the 21st International Conference on Machine Learning*.

Georghiades, A., Belhumeur, P., & Kriegman, D. (2001). From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Trans. Pattern Anal. Mach. Intelligence*, *23*.

Griffiths, T., & Ghahramani, Z. (2005). Infinite latent feature models and the Indian buffet process. *TR 2005-001, Gatsby Computational Neuroscience Unit*.

Ishwaran, H., & James, L. F. (2001). Gibbs sampling methods for stick breaking priors. *Journal of the American Statistical Association*, *96*, 161–173.

Knowles, D., & Ghahramani, Z. (2007). Infinite Sparse Factor Analysis and Infinite Independent Components Analysis. *Lecture Notes in Computer Science*, *4666*, 381.

Meeds, E., Ghahramani, Z., Neal, R. M., & Roweis, S. T. (2007). Modeling dyadic data with binary latent factors. *Advances in Neural Information Processing Systems 19*.

Teh, Y. W., Gorur, D., & Ghahramani, Z. (2007). Stick-breaking construction for the Indian buffet process. *Proceedings of the 11th Conference on Artificial Intelligence and Statistic*.

Thibaux, R., & Jordan, M. (2007). Hierarchical beta processes and the indian buffet process. *Proceedings of the International Conference on Artificial Intelligence and Statistics*.

Wainwright, M. J., & Jordan, M. I. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, *1*, 1–305.

Winther, O. (2000). Gaussian processes for classification: Mean field algorithms. *Neural Computation*, *12*.

Wood, F., & Griffiths, T. L. (2007). Particle filtering for nonparametric Bayesian matrix factorization. In *Advances in Neural Information Processing Systems 19*.